mathematics

MDPI

Article

Deep-Learning-Based Remaining Useful Life Prediction Based on a Multi-Scale Dilated Convolution Network

Feiyue Deng [1,2], Yan Bi [2], Yongqiang Liu [1,*] and Shaopu Yang [1]

[1] State Key Laboratory of Mechanical Behavior and System Safety of Traffic Engineering Structures, Shijiazhuang Tiedao University, Shijiazhuang 050043, China; dengfy@stdu.edu.cn (F.D.); yangsp@stdu.edu.cn (S.Y.)
[2] School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China; biyan2015halfwave@163.com
* Correspondence: liuyq@stdu.edu.cn; Tel.: +86-311-87936742

Abstract: Remaining useful life (RUL) prediction of key components is an important influencing factor in making accurate maintenance decisions for mechanical systems. With the rapid development of deep learning (DL) techniques, the research on RUL prediction based on the data-driven model is increasingly widespread. Compared with the conventional convolution neural networks (CNNs), the multi-scale CNNs can extract different-scale feature information, which exhibits a better performance in the RUL prediction. However, the existing multi-scale CNNs employ multiple convolution kernels with different sizes to construct the network framework. There are two main shortcomings of this approach: (1) the convolution operation based on multiple size convolution kernels requires enormous computation and has a low operational efficiency, which severely restricts its application in practical engineering. (2) The convolutional layer with a large size convolution kernel needs a mass of weight parameters, leading to a dramatic increase in the network training time and making it prone to overfitting in the case of small datasets. To address the above issues, a multi-scale dilated convolution network (MsDCN) is proposed for RUL prediction in this article. The MsDCN adopts a new multi-scale dilation convolution fusion unit (MsDCFU), in which the multi-scale network framework is composed of convolution operations with different dilated factors. This effectively expands the range of receptive field (RF) for the convolution kernel without an additional computational burden. Moreover, the MsDCFU employs the depthwise separable convolution (DSC) to further improve the operational efficiency of the prognostics model. Finally, the proposed method was validated with the accelerated degradation test data of rolling element bearings (REBs). The experimental results demonstrate that the proposed MSDCN has a higher RUL prediction accuracy compared to some typical CNNs and better operational efficiency than the existing multi-scale CNNs based on different convolution kernel sizes.

Keywords: remaining useful life; deep learning; rolling element bearing; multi-scale feature fusion; dilated convolution

check for updates

Citation: Deng, F.; Bi, Y.; Liu, Y.; Yang, S. Deep-Learning-Based Remaining Useful Life Prediction Based on a Multi-Scale Dilated Convolution Network. *Mathematics* **2021**, *9*, 3035. https://doi.org/10.3390/math9233035

Academic Editors: José Rodellar, Francesc Pozo and Yolanda Vidal

Received: 24 October 2021
Accepted: 23 November 2021
Published: 26 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the development of industrial science and engineering maintenance, the level of automation and complexity of mechanical equipment (ME) has been increasing. Under the comprehensive interaction of internal and external factors, the performance and health status of mechanical components would inevitably degrade. When the degradation reaches a certain extent, equipment will suffer serious failure, resulting in irreparable economic losses and a waste of resources [1].

In the context of the continuous development of Industry 4.0 and the Internet of Things, prognostic and health management (PHM) technology has been widely used for fault diagnosis studies of ME key parts. Rolling element bearing (REB) is the one of principal factor determining the state of health of rotating machinery. The remaining useful

life (RUL) prediction can identify damage and defects of the REB as early as possible to avoid equipment damage and personnel casualties [2]. Therefore, bearing RUL prediction is of great significance.

Generally, the RUL prediction paradigm includes three kinds of methods: model-based, experience-based, and data-driven approaches [3]. The use of general model-based or experience-based approaches depends on a deep understanding of system-failure physics and requires complete knowledge of the dynamics to construct model equations. Due to the complexity of the mechanical system structure, it is difficult to be accurately applied in the practical engineering system. In contrast, the data-driven approach is primarily being prepared on the basis of sensor data, and it requires less knowledge of inherent system failure mechanisms. Failure regularity of the REB can be identified more accurately by data analysis only. This has led to many data-driven approaches being put forward in recent years, such as artificial neural network (NN) [4], support vector machine (SVM) [5], k-nearest neighbor (KNN) [6], hidden Markov modes [7], etc. Liu et al. [8] proposed an RUL prediction model based on multiple health state assessments and then predicted the RUL of REB using the SVM. Zheng [9] extracted the useful features by the empirical mode decomposition (EMD) and estimated the bearing's current state using the RUL prediction model based on the KNN classifier. Ail et al. [10] adopted the ANN to construct an RUL prediction model for pump bearings, which has good properties for analyzing time-domain signals. Although the above studies had a certain degree of successful RUL prediction, most of them need complicated signal processing techniques and some prior knowledge to extract feature information. Moreover, these models are all shallow network architectures. In realistic applications, they are inapplicable to automatic analysis and feature extraction for big data due to requiring a large amount of prior knowledge.

Deep learning (DL) technology can construct a deeper nonlinear network structure, achieving the approximation of complicated functions and characterization of the analyzed data. Therefore, DL is increasingly attractive in the data-driven prognostics model. A deep neural network (DNN) based on DL can automatically mine feature information from input raw data by stacking multilayer neural networks, which overcomes the dependence on signal processing technology and expert knowledge, and has obvious advantages in analyzing big data [11,12]. In recent years, a growing amount of studies have been dedicated to constructing a DNN-based RUL prediction model, leading to the emergence of prognostics models utilizing different types of networks. Deep belief network (DBN), auto-encoder network (AEN), recurrent neural network (RNN), and convolution neural network (CNN) are mainstream architectures in DL. Deutsch et al. [13] proposed a DBN-based prognostics model for RUL prediction of rotating components, which was validated through gear tests and bearing run-to-failure tests. DBN and AEN are time-consuming because their training processes contain too many weight parameters. Long short-term memory (LSTM), as an improvement of the RNN, is good at extracting the sequential nature of the temporal signal. Hinchi et al. [14] presented an end-to-end framework for REB RUL estimation based on convolutional and LSTM recurrent units. Chu et al. [15] constructed an integrated network model with CNN and LSTM to extract signal features and estimate the RUL of bearing. The LSTM is based on a serial computing system, which leads to low operational efficiency and limits its application in engineering.

Among the above DNNs, CNNs have shown a remarkable ability in extracting features information from the condition monitoring data. CNN can effectively reduce the computational burden in the training process because it utilizes a local receptive field and weight-sharing strategies to decrease the number of weight parameters [16]. Ren et al. [17] adopted a CNN-based network model with 13 convolution layers to predict REB RUL, in which the spectrum-principal-energy vector was used to extract signal eigenvectors to input the network. Li et al. [18] combined short-time Fourier transform (STFT) and CNN to acquire the RUL of REB, where raw vibration signals are pre-processed with the STFT to extract time-frequency features and used as the input eigenvectors. Although

CNN-based prognostics models exhibit promising prognostic performance, there remain the following shortcomings:

(1) Traditional CNNs have a fixed convolution kernel size for each convolution layer in the network. The ability to extract features for CNN would decline significantly under a strong noise environment because the specific scale convolution kernel can only learn feature information of the corresponding scale.
(2) Numerous multi-scale CNN models [12,16,19] have been proposed recently to extract different-scale feature information. Even though these models improve the learning ability of the network, the existing multi-scale networks' frameworks are all based on different sizes of convolution kernels. A multi-scale convolution framework based on multiple size convolution kernels requires more convolution operations, much greater computation, and a mass of weight parameters. A low operation speed severely restricts its application in engineering applications.
(3) A complex network architecture that contains too many convolution layers is not appropriate for prognostic prediction. The increase in network depth does not further improve the network performance effectively [20], and the network training time increases dramatically. In addition, too deep of a network structure can easily result in overfitting in the case of small datasets.

To tackle the above problems, a one-dimensional (1D) end-to-end multi-scale deep CNN model for RUL prediction was proposed in this study, and the sketch map of the proposed method is shown in Figure 1. The monitoring 1D data can be directly fed into the proposed network without any pre-processing operations. It does not require complex signal processing techniques, increasing the commodity and applicability of the proposed model largely. The proposed multi-scale dilated convolution network (MsDCN) adopts a novel multi-scale dilated convolution fusion unit (MsDCFU) instead of the traditional multiple convolution kernels with different sizes. The MsDCFU can extract different-scale feature information using diverse size dilation factors to expand the receptive field of the convolution kernel, almost without additional computational burden. In addition, depthwise separable convolution (DSC) was employed to replace the standard convolution in the proposed model to further reduce the network parameters and the computation cost. The rest of the article is structured as follows. The standard convolution, dilation convolution, and DSC theory are described in Section 2. The proposed MsDCFU and multi-scale model framework are described in detail in Section 3. Experiment and comparison analyses are illustrated in Section 4. Conclusions are composed in Section 5.

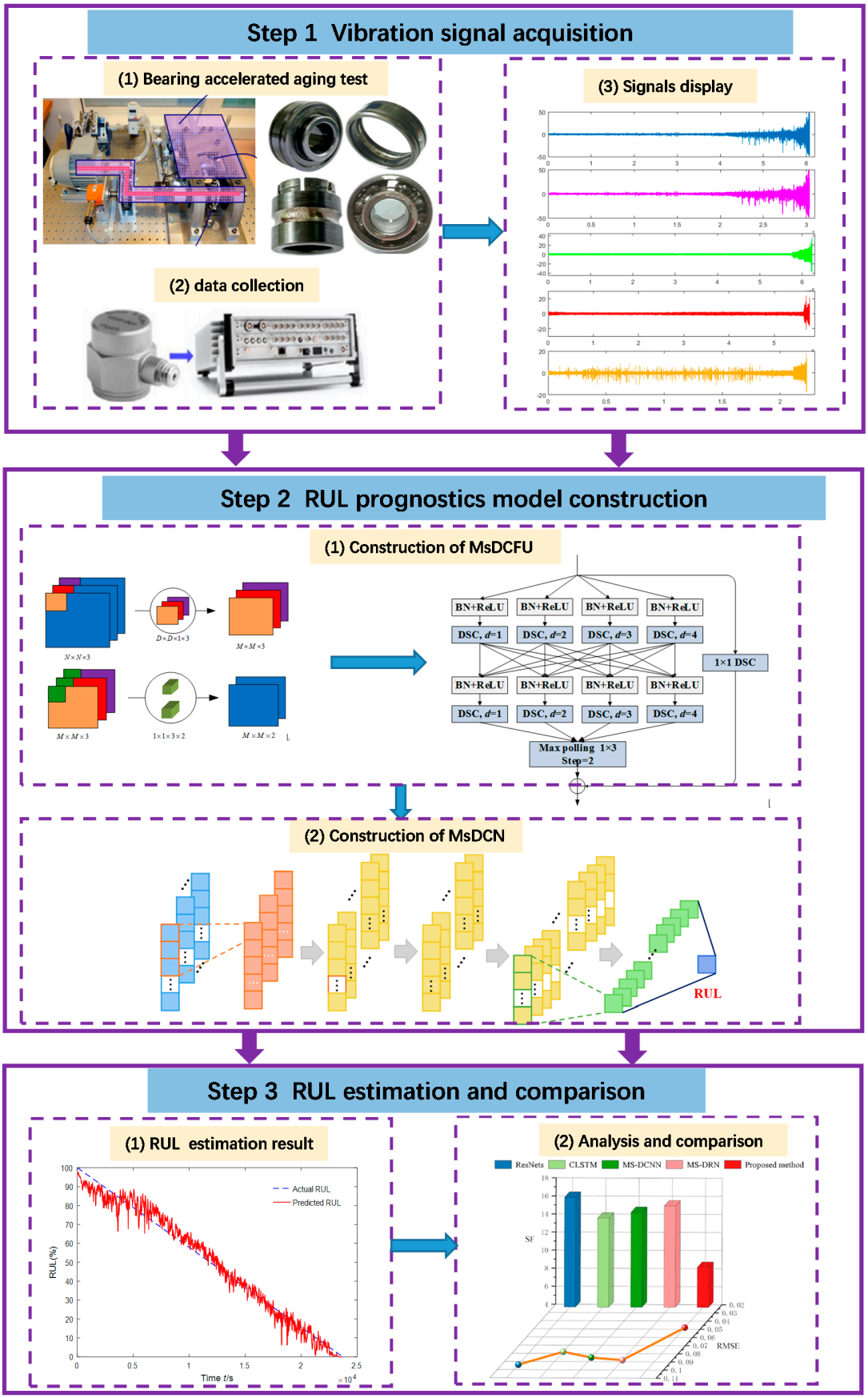


Figure 1. Sketch map of the proposed method.

2. Preliminaries

2.1. Convolutional Neural Network (CNN)

CNN, as a kind of feedforward neural network, mainly consists of multiple convolutional layers, pooling layers, and fully connected (FC) layers. In addition, the batch normalization (BN) layer and the activation layer are also important parts of CNN. CNN adopts the strategy of the local receptive field and weight sharing, with high training efficiency and fewer network parameters, thus making it widely used in image recognition, speech processing, and other fields [21].

The convolution layer is the most important part of the CNN, which extracts the learned features to the next layer by performing local convolutional operations on the output features of the network in the previous layer. This process can be described by this formula:

$$y = f(W * x + b) \tag{1}$$

where x is the input of the convolution layer, y denotes the convolution layer output result, W is the weight parameter, and b is the bias of the convolution kernel. $f()$ denotes the activation function, which is used to enhance the nonlinear expression ability of learned features. The rectified linear unit (ReLU) is the most widely used nonlinear mapping activation function. It not only accelerates the convergence of the model but also makes the network easier to optimize in the process of back-propagation learning.

The pooling layer is usually added after the convolution layer, which can compress the feature map size, reduce the dimension of spatial features, and simplify the network's complexity. The most common pooling methods contain average pooling, max pooling, L2-norm pooling, and global average pooling (GAP). GAP is used to calculate the mean value of each channel feature map, which is often used before the FC layer. The BN layer is arranged between the convolution layer and the activation layer to reduce the shift of internal covariance and to accelerate the training process of multilayered networks. The FC layer is located at the end of the entire CNN and acts as a classifier by using the softmax function to classify the learned feature information with minimum cross-entropy loss as the objective function.

2.2. Dilated Convolution

Traditional CNNs normally use standard convolution operations, in which the dilation factor is fixed, usually $d = 1$, which denotes that there is no interval between the convolution kernel and the elements of the feature map during the convolution process. The range of the receptive field in the convolution operation for the dilation factor $d = 1$ corresponds to the size of the convolution kernel. The convolution operation for the standard convolution with dilation factor $d = 1$ is illustrated in Figure 2a. From the figure, the range of receptive field is 3×3, which is the same size as the convolution kernel.

Dilated convolution is defined as a convolution operation with different degrees of spacing between the elements of the feature map when the dilation factor is set to $d > 1$. Yu et al. [22] proposed a dilated convolution neural network for image segmentation and validated that the dilated convolution can increase the receptive field of CNN without introducing additional parameters. Zhang et al. [23] constructed a compressed dense block, where dilated convolution is introduced to obtain a large receptive field without reducing the loss of feature information. Wei et al. [24] designed a generic classification network equipped with dilated convolutional blocks of different dilation factors. Generally, the receptive field of the dilation convolution is larger than the standard convolution ($d = 1$). The convolution operation of the dilated convolution with dilation factor $d = 2$ is displayed in Figure 2b. It can be seen that the range of the receptive field is 5×5. Comparing the two figures, it can be found that although convolution kernel size is the same in the standard convolution and dilated convolution, the latter has a larger receptive field range during the convolution process, and thus it can extract larger-scale feature information. Different scales of dilation factors can extract feature information at different scales, and with increasing dilation factors, the range of the corresponding receptive field range increases significantly.

It should be further noted that the dilated convolution utilizes an increase in the dilation factor to increase the range of the receptive field. Therefore, it does not bring an increase in the number of extra parameters and computation in the convolution operation compared to the standard convolution operation.

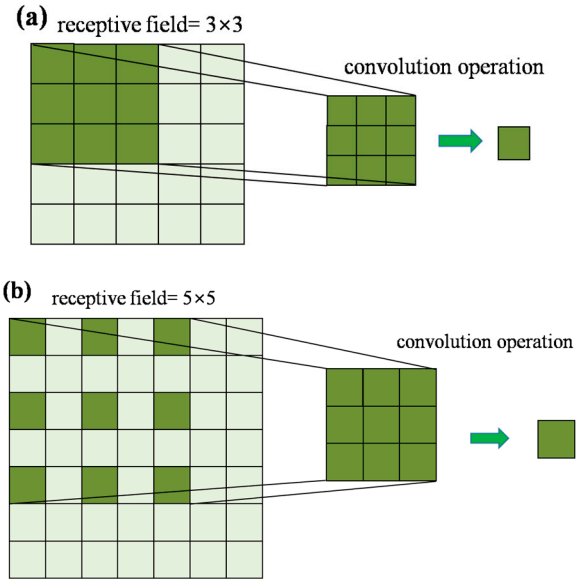


Figure 2. Convolution operation with different dilation factors: (**a**) $d = 1$ and (**b**) $d = 2$.

2.3. Depthwise Separable Convolution (DSC)

During the standard convolution operation, the convolution kernels are convoluted with the feature map of each channel separately, and the output features are obtained by linear superposition of the convolution results. The standard convolution process is shown in Figure 3, in which the input feature map dimension is $N \times N \times 3$, the output feature map dimension is $M \times M \times 2$, and the convolution kernel dimension is set up as $D \times D \times 3 \times 2$. It can be seen from the figure that both the number of channels for the input features and the convolution are taken into account simultaneously in the standard convolution process. The number of parameters in the standard convolution operation is as follows:

$$Q_{standard} = D \times D \times 3 \times 2 \tag{2}$$

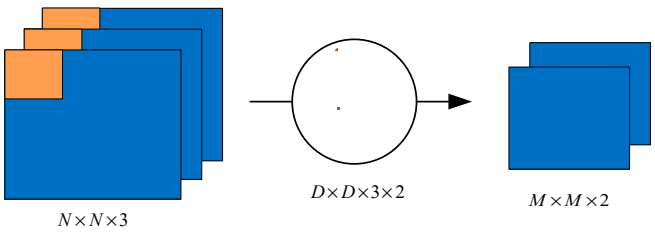


Figure 3. Standard convolution operation.

The DSC operation includes a sequence of depthwise convolution (DWConv) and 1×1 pointwise convolution (PWConv) [25], which are shown in Figure 4a,b, respectively. The number of convolution kernels is the same as the number of input feature channels in the DWConv, and the convolution kernels only perform the convolution operation with the corresponding channel input features; thus, the number of output feature channels is the same as the number of input feature channels. PWConv has the same convolution process as the standard convolution operation but uses a size of 1×1 convolution operation. Compared to standard convolution, DSC can greatly reduce the size of the network and

learn representative features faster and more accurately. More detail about the DSC operation process can be found in ref. [26]. The convolution area is considered first in the DSC, followed by the channels of the input features through which the separation of channels and convolution areas is achieved. The number of parameters in the DSC operation is:

$$Q_{DSC} = D \times D \times 3 + 3 \times 1 \tag{3}$$

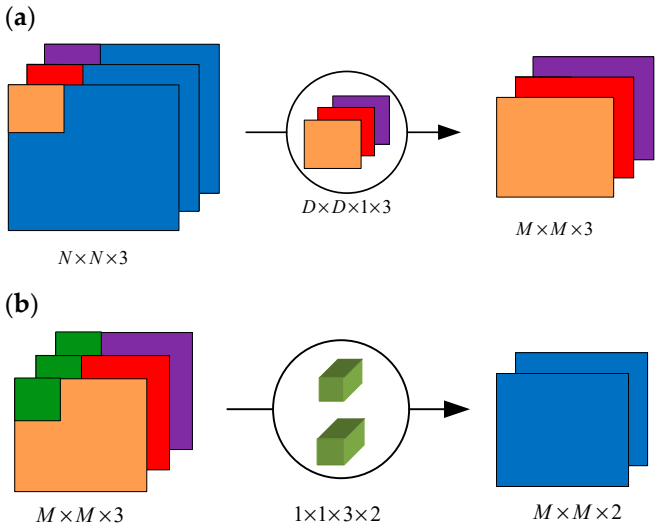


Figure 4. Depthwise separable convolution: (**a**) DWCnov and (**b**) PWConv.

Comparing Equations (2) and (3), the number of parameters for the DSC is much less than that of standard convolution. When the number of convolution kernels or the number of feature map channels increases dramatically, the reduction in the number of parameters for the DSC will be considerable. Moreover, Howard et al. [27] further demonstrated that DSC also requires much less computation than the standard convolution, and its operation is more highly efficient.

3. Multi-Scale Dilated Convolution Network (MsDCN)

3.1. Multi-Scale Dilated Convolution Fusion Unit (MsDCFU)

The vibration signal of rotating machinery is a typical nonlinear and non-smooth signal. The signal components are very complex, and the feature information is distributed in separate frequency bands, so extracting multi-scale feature information can better predict the evolution law of mechanical equipment service state. Jia et al. [28] confirmed through their study that the convolution layer of CNN is equivalent to a band-pass filter, and the feature information of different frequency bands in the signal can be extracted by setting the relevant parameters of the convolution layer. The receptive field range of the convolution kernel varies, and the scale of the feature information extracted in the convolution operation varies. A larger receptive field can obtain larger-scale feature information in the signal, while a smaller receptive field reflects detailed feature information. As a consequence, some multi-scale CNN models based on different sizes of convolution kernels have been proposed recently [16,18,19] to enhance the CNNs' feature information mining capability. Deng et al. [16] designed a multi-scale feature fusion block based on four different convolution sizes 1×1, 3×3, 5×5 and 7×7, which can learn different-scale feature information. Li et al. [19] also proposed a multi-scale block consisting of three different convolutional kernel sizes of 10×1, 15×1, and 20×1, which can learn the mapping relationship more accurately.

Both changing the size of the convolution kernel and the dilation factor can increase the receptive field range in the convolution operation, but the former will lead to an increase in the number of network parameters and a decrease in the calculating efficiency, while

the latter can effectively expand the receptive range with little change in the amount of the convolution operation. Inspired by this, a new multi-scale CNN prognostic model whose convolution layer has different ranges of receptive fields is proposed to extract the feature information of the temporal signal with different scales, achieving the purpose of improving the prediction performance of the network. Unlike most existing studies that construct multi-scale network models by setting different convolution kernel sizes, this study designed a MsDCFU by fusing multiple convolution operations with different dilation factors, whose structure is depicted in Figure 5.

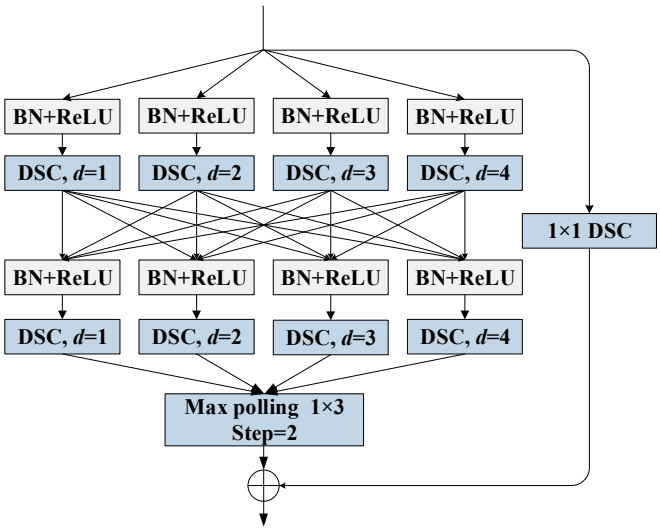


Figure 5. Architecture of the MsDCFU.

The proposed method adopts a residual block connection structure, including a cross-layer identity mapping connection and a main connection consisting of multiple multi-scale dilated convolution operations. The residual block structure effectively solves the problem of gradient vanishing when the depth of the network increases, having a higher network training speed and easier model parameter optimization. The proposed method achieves the extraction and fusion of multi-scale feature information within the residual block, taking into account the advantages of cross-layer identify mapping and multi-scale feature extraction. In the proposed MsDCFU, the input data are first carried out with four different scales of d = 1, 2, 3, and 4 dilated convolution operations to extract feature information with different receptive field ranges in the temporal signal. Different-scale feature information is fused into a new feature map through the concatenate operation, and then the multi-scale dilated convolution with different dilation factors and concatenating operations are performed again. Finally, a max-pooling operation with step = 2 and size = 1 × 5 is used to reduce the feature map dimension and further extract more essential feature information. To improve the generalization ability and the convergence speed of the network, both BN and ReLU are employed before the DSC operation. In addition, a 1 × 1 DSC is adopted in the identity mapping connection of the residual block to ensure that feature dimensions at the input and output of the residual block are the same. It should be pointed out that the size of the convolution kernels for the DSC in the proposed method is 1 × 5. A thorough comparative analysis of the influence of the convolution kernel size on the RUL prediction performance of the network model is conducted in the experimental section.

3.2. Architecture of MsDCN

The structure of the proposed MsDCN is illustrated in Figure 6. The input of the network is a 1D vibration signal. The first layer of the network is a standard convolution layer, and, after that, three MsDCFUs are connected. The number of the MsDCFU affects the prediction performance and operational efficiency of the model. Too many MsDCFUs

will lead to a deeper network layer and poor training results; too few MsDCFUs will result in deficient feature-learning capabilities. Therefore, we made a further comparative analysis on the influence of the MsDCFU number in the experimental section. It should be noted that the max pooling layer is removed in the third MsDCFU in order not to make the input data dimension too small. Then, the feature map information is processed by the GAP and dropout layer (dropout rate = 0.5), alleviating the effect of the over-fitting problem. Finally, the proposed model applies the FC layer to output the prediction result of the REB RUL. The detailed network parameters for the proposed model are presented in Table 1. In addition, the L2 regularization technique is used to alleviate overfitting of the network model, and the Adam algorithm is employed to adaptively optimize the network parameters.

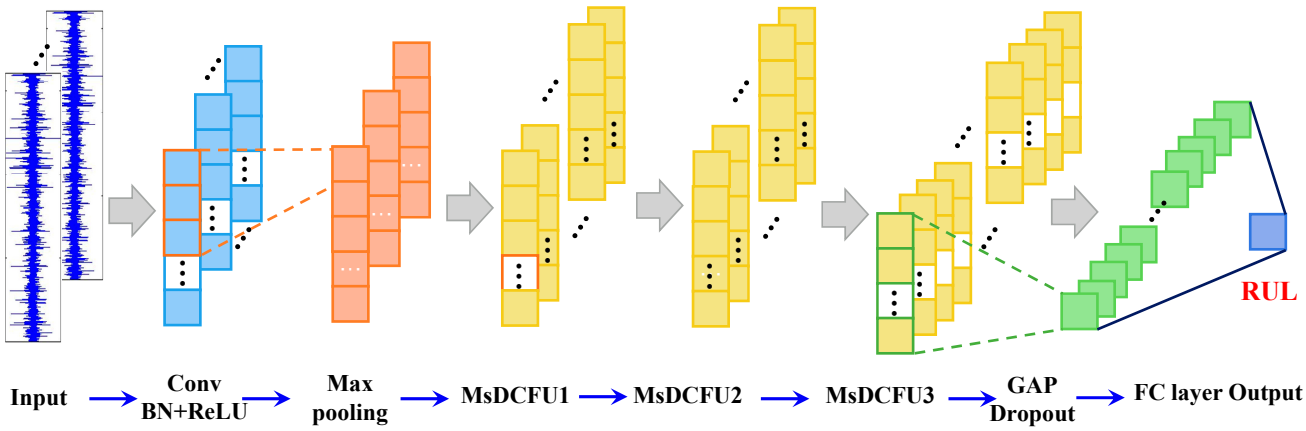


Figure 6. Architecture of the proposed MsDCN.

Table 1. Parameter setting of network.

Layer Type	Kernel Size/Stride	Dilation Rate *d*	Kernels Number/Sum	Padding
Standard convolution	Conv [1 × 49]/8	1	64/64	Yes
Maxpooling	[1 × 5]/1			Yes
MDCFU1	DSC [1 × 5]/1	1/2/3/4	16/64	Yes
	DSC [1 × 5]/1	1/2/3/4	16/64	Yes
	Maxpooling [1 × 3]/2			Yes
MDCFU2	DSC [1 × 5]/1	1/2/3/4	16/64	Yes
	DSC [1 × 5]/1	1/2/3/4	16/64	Yes
	Maxpooling [1 × 5]/2			Yes
MDCFU3	DSC [1 × 5]/1	1/2/3/4	32/128	Yes
	DSC [1 × 5]/1	1/2/3/4	32/128	Yes

3.3. Sample Signal Processing and Label Generation

Before being input to the prognostics model, the signal samples need to be normalized and time-window embedded. The data signal is normalized as follows:

$$x_{norm} = \frac{x - \mu}{\delta} \tag{4}$$

where x_{norm} is the normalized signal, and μ and δ are the mean value and standard deviation of the original signal, respectively.

Wang et al. [29] demonstrated that the prediction performance of data-driven prognostics models based on DNNs can be improved drastically by embedding window information into the input data in the RUL prediction. Therefore, a fixed-size time window is incorporated into the model input data according to the time window embedding strategy. A fixed-size time window is selected as a time step to concatenate the measured signal after

normalization into a high-dimensional vector, and then the vector is fed to the model as an input. Assuming that time step $t = 3$, the process of time window embedding is displayed in Figure 7. The input vector acquired by the time step embedding can be illustrated as follows:

$$x_{input}^{j} = (x_{norm}^{j}, x_{norm}^{j+1}, \cdots, x_{norm}^{j+t-1}) \tag{5}$$

where j is the sequence number of input data, and t is the time step. In this study, many comparative analyses of the model prognostics performance for different time step sizes were performed according to the literature [25], and they finally took $t = 5$.

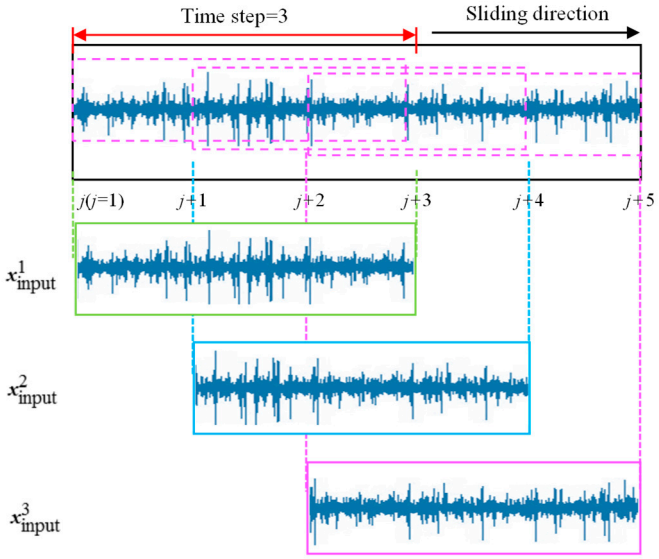


Figure 7. Illustration of time-window embedding.

There is considerable randomness in the accelerated aging test of bearings, and the actual lifetime range is very wide. Hence, it is necessary to normalize the labels of bearing whole lifetime samples to be within the range of [0, 1], then using the normalized RUL results as the output labels of the network model. The RUL result normalization process is formulated as:

$$y_{i,j} = \frac{N - j}{N - 1} \tag{6}$$

where $i = 1, 2, \ldots, 7$, N is the total number of samples in the i-th whole life bearing test, and j represents the sequence number of the i-th whole life bearing test. The discrepancies between different RUL values of each bearing are minimized by normalizing the lifetime labels of different bearings, which facilitates the network model gradient descent optimization solution.

4. Experimental Verification

4.1. Data Descriptions

The experimental data are from the PHM 2012 Challenge dataset, which was collected at the accelerated aging test rig PRONOSTIA [30]. The structure of the test rig is shown in Figure 8. A synchronous motor and an assembly of two pulleys were applied for the tested bearing speed regulating. The load generated by the pneumatic jack is transmitted in the radial direction of the bearing through the force transmission device. Two accelerometers of type 3035B DYTRAN were mounted horizontally and vertically on the tested bearing to acquire the whole-life vibration signals. The sampling frequency of the data was 25.6 kHz, and the data was recorded every 10s, with each acquisition time being 0.1 s. There were three working conditions for the tested bearing in the test, and the detailed information of the experiments can be found in ref. [31]. In this study, we mainly analyzed the bearing degradation process when the load was 4000 N and the rotating speed was 1800 r/min.

In this condition, seven whole-life bearing tests were conducted, which are divided into bearing 1_1 to bearing 1_7. Bearing 1_3 data were chosen to be the testing set, and the rest composed the training set.

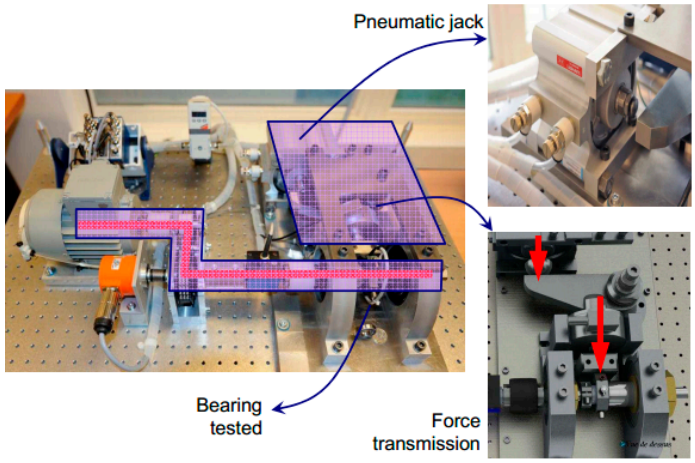


Figure 8. Bearing-accelerated degradation test rig.

Since the tested bearing was subjected to the radial load, the radial vibration signal collected by the accelerometer was taken as input to the network model. The time-domain waveform of the whole life-cycle for the tested bearing 1_3 is shown in Figure 9. It can be seen from the figure that the waveform amplitudes of the radial vibration signals are small and relatively smooth in the early test phase, which indicates that the bearing is operating in a healthy state. As the experiments progressed, the amplitude of the collected vibration signal waveform gradually increased, indicating that the normal operating condition of the bearing had begun to change and that a relatively minor failure has occurred. At the end of the experimental procedure, the amplitude of the signal waveform increased sharply, which indicates that the failure of the bearing had become very evident and that the bearing had failed.

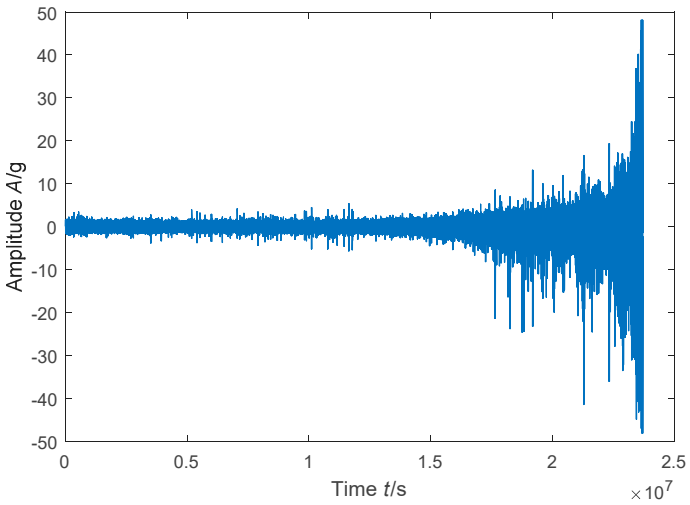


Figure 9. Whole life-cycle time domain waveform of bearing 1_3.

4.2. Experimental Study

The scoring function (SF) and the root mean square error (RMSE) are two commonly used evaluation indexes in the field of PHM and were chosen to quantitatively evaluate the prediction performance in the experiment. SF is from the 2008 Prognostics and Health Management Data Challenge. The smaller the absolute error value is, the lower the SF

value is, and the RMSE has the same characteristic. The experimental test was repeated 10 times to reduce the impact of randomness.

In this case study, two main hyperparameters that may affect the prediction performance of the proposed model were first investigated in detail: they were the number of the MsDCFUs and the size of the convolution kernels. Then, the proposed MsDCN network was compared with five state-of-the-art prognostics models to show its accuracy and rapidity. As analyzed in the above section, more MsDCFUs will result in a deeper network structure, which aggravates the calculation burden. Moreover, the limited training dataset may lead to over-fitting during network training as the network depth increases. When the convolution kernel size is taken as 1×5 by default, the number of MsDCFUs is set to be 2, 3, 4, and 5 respectively. The proposed models based on different MsDCFUs numbers were applied to the RUL estimation of bearing. The 2-D histograms of the SF values and RMSE values of the tested bearing for different numbers of MsDCFUs are displayed in Figure 10. It can be seen that when the number of MsDCFUs was set to 3, both the SF value and the RMSE value were minimal. Then, the training time and model parameters based on different numbers of MsDCFUs were calculated and presented in Table 2. It is clear that with the increase of in the number of MsDCFUs, the model parameters and training time increased significantly. To achieve a good trade-off between the calculating cost and RUL prediction accuracy, the number of the MsDCFUs was finally chosen to be 3.

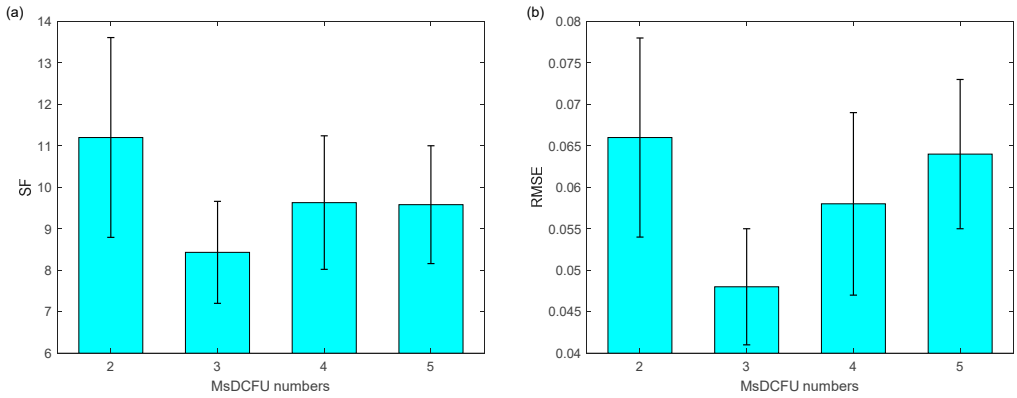


Figure 10. Evaluation indexes based on different MsDCFU numbers: (**a**) RMSE and (**b**) SF.

Table 2. Comparison of model parameters and training time with different MsDCFU numbers.

MsDCFU Number	2	3	4	5
Training time/s	23,374	23,886	24,979	25,958
Total model parameters	57,537	73,153	112,321	168,129

The convolution kernel size is another key hyperparameter in the proposed MsDCFU, and it directly affects the dimension of features extracted in the DSC operation. For investigating this influence, the DSC operations with different kernel sizes in the proposed MsDCFU were applied to estimate the RUL prediction of bearing. When the number of MsDCFUs was taken as 3 by default, the convolution kernel size was set to be 1×3, 1×5, 1×7, and 1×9, respectively. Figure 11 shows the 2-D histograms of the SF values and RMSE values for the RUL estimation of the tested bearing, and the corresponding training time and model parameters are given in Table 3. It can be observed that the SF value and the RMSE value were largest when the convolution kernel size was set to 1×3, which indicates that the accuracy of the RUL estimation was relatively poor. The accuracy of the RUL prediction results was closer for the other three convolution sizes. As the size of the convolution kernel increased, the model was significantly more computationally intensive. Therefore, it can be observed in Table 3 that the model training time and the number of parameters increased significantly with the increase in the convolution kernel size, and the

network operation efficiency decreased markedly. By the trade-off between accuracy and efficiency, the convolution kernel size in the MSDCFU was finally selected as 1×5.

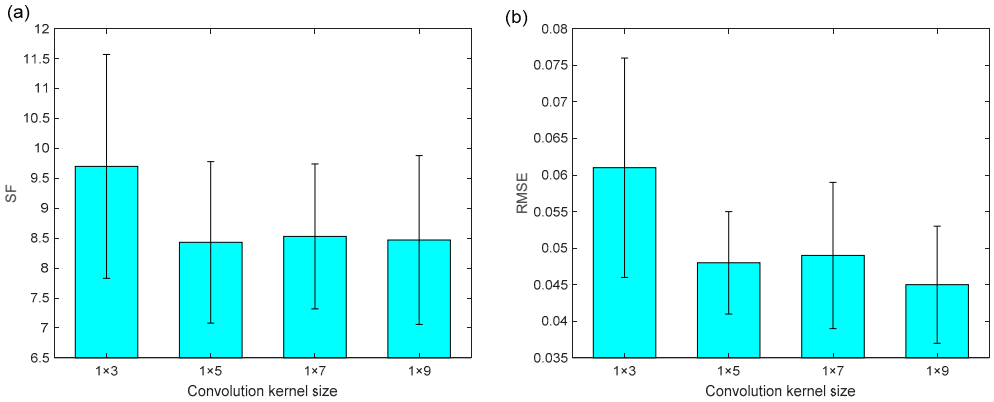


Figure 11. Evaluation indexes based on different convolution kernel sizes: (**a**) RMSE and (**b**) SF.

Table 3. Comparison of model parameters and training time with different convolution kernel sizes.

Convolution Kernel Size	1×3	1×5	1×7	1×9
Training time/s	23,527	23,886	24,188	24,490
Total model parameters	69,568	73,153	76,737	80,321

When the model hyperparameters were determined, the proposed network model was used to predict the RUL of bearing 1_3, and the RUL prediction result is shown in Figure 12. As can be seen from the figure, the predicted value of RUL fluctuates slightly with the actual RUL, and the RUL predicted result accurately reflects the degradation process of the lifetime for the bearing 1_3.

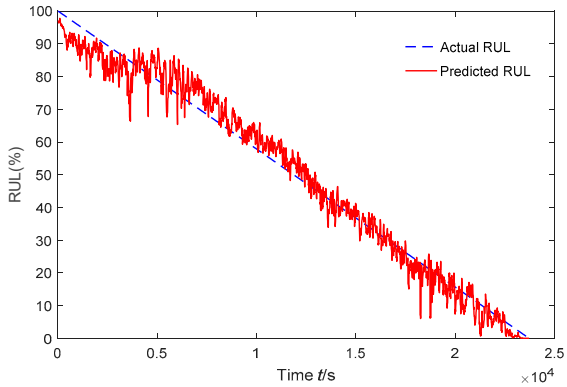


Figure 12. RUL prediction result of bearing 1_3 using the proposed method.

Two existing common prognostics models ResNets [32] and convolutional LSTM (CLSTM) [33], and two state-of-the-art multi-scale CNN prognostics models, including multi-scale deep convolutional neural network (MS-DCNN) [19] and multi-scale deep residual network (MS-DRN) [16], were utilized to estimate the bearing RUL for the comparison analysis. Figure 13 displays the RUL prediction results of the above four comparison methods, and Figure 14 summarizes their evaluation indexes for RUL estimation. Figure 13 displays the RUL prediction results of the above four comparison methods, and Figure 14 presents specific SF values and RMSE values of the RUL prediction results. Comparing Figures 12 and 13, it was found that the RUL prediction values by the proposed method agree well with the actual RUL values, which can achieve relatively accurate bearing RUL prediction. Among the four comparison methods, the prediction

values of ResNets deviate the most from the actual RUL values, which indicates that the ResNets approach is difficult to achieve an accurate RUL prediction result. The remaining three methods were relatively accurate in the early stages of RUL prediction, while more significant deviations were observed in the later stages. Among the above five methods, the SF value and the RMSE value of the proposed method were the lowest. It can be concluded that the RUL prediction accuracy of the proposed method was superior to the four comparison approaches.

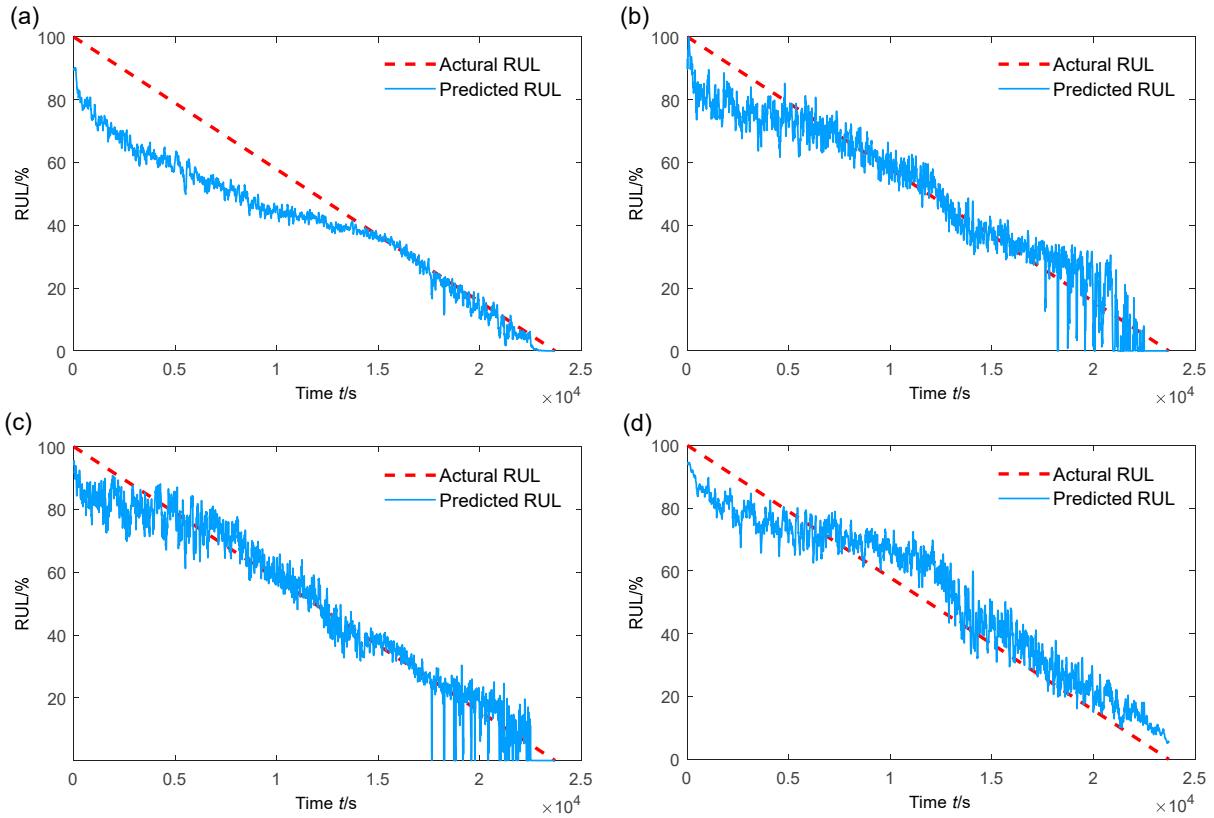
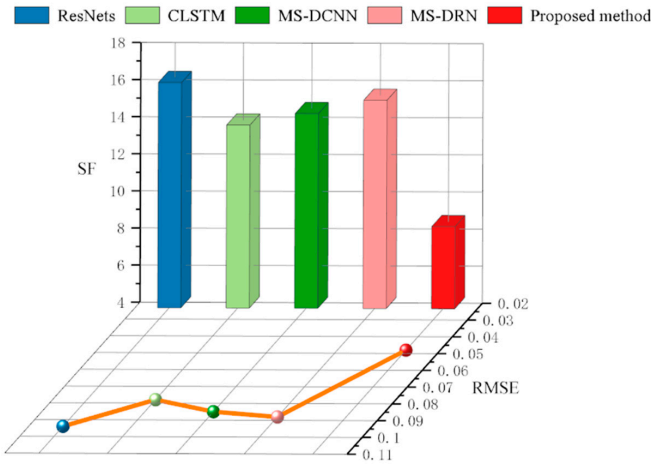


Figure 13. RUL prediction results of bearing1_3 for four prognostics approaches: (**a**) ResNets; (**b**) CLSTM; (**c**) MS-DCNN; and (**d**) MS-DRN.



	ResNets	CLSTM	MS-DCNN	MS-DRN	Proposed method
SF	16.15±2.54	13.87±2.10	14.50±1.90	15.23±0.22	8.43±0.18
RMSE	0.094±0.014	0.078±0.011	0.085±0.013	0.088±0.012	0.048±0.005

Figure 14. Evaluation indexes of different methods for the RUL prediction.

To further show the superiority of the proposed method in operation efficiency, Table 4 exhibits the total model parameters and operation times of MS-DCNN, MS-DRN, and the proposed method. It can be seen that the total model parameters of the proposed method were reduced by 96.1% and 36.6% compared to the MS-DCNN method and the MS-DRN method, respectively. The training time and the testing time of the proposed method were greatly reduced, while the operation efficiency was significantly improved.

Table 4. Prediction performance comparison of the proposed method and the other two multi-scale CNN-based prognostics models.

Method	Total Model Parameters	Operation Time/s	
		Training	Testing
MS-DCNN	1,864,833	56,574	13
MS-DRN	113,889	46,748	12
Proposed method	73,153	23,886	7

5. Conclusions

Compared with traditional CNN, multi-scale CNN can extract feature information of different scales. Therefore, the multi-scale CNN-based prognostics model has greatly improved learning capability and has a better performance in remaining useful life (RUL) prediction. However, the existing multi-scale CNNs are all based on different sizes of convolution kernels, which need more convolution operations and a mass of weight parameters. This leads to low computational efficiency and is difficult to be widely used in engineering applications. To address the issue, a one-dimensional (1D) end-to-end multi-scale dilated convolution network (MsDCN) for RUL prediction of REB was presented in this study. The innovations of the proposed prognostics model are summarized as follows. (1) Different from the traditional multi-scale CNN framework, a new multi-scale dilation convolution fusion unit (MsDCFU) was adopted in the proposed method. The MsDCFU can extract multi-scale feature information using different dilated factors and can expand the receptive field (RF) of the convolution kernel with no additional computational burden. (2) The standard convolution was replaced by the depthwise separable convolution (DSC) in the MsDCFU, which further reduces the total number of model parameters and improves the efficiency of the operations. The proposed method was experimentally validated by using the PHM 2012 Challenge dataset for the accelerated degradation of REBs. Some state-of-the-art prognostics models were also analyzed for comparison with the proposed method in this studdy. The experimental results indicate that the proposed method has a higher RUL prediction accuracy compared to typical CNNs and is strikingly more efficient than the existing multi-scale CNNs based on different convolution kernel sizes.

Although a good bearing RUL prediction result was achieved by the proposed MsDCN approach, there are still a few shortcomings in its application, including the issue of adaptive optimization of network hyperparameters and the quantification of uncertainty in the RUL prediction result. To adaptively optimize the network hyperparameters, a promising work is to design a new loss objective function during the process of network model training according to the characteristics of RUL prediction. Simultaneously, this can be combined with some adaptive optimization algorithms, such as the particle swarm optimization and the whale optimization algorithm, to automatically determine reasonable network hyperparameters. As for quantification of uncertainty in the RUL prediction, on the one hand, multi-sensor data are used to improve the robustness of RUL prediction. Different sensor data contain different degrees of degradation information, some of which may be sensitive to the degradation of the machine, but others may not. Analyzing multiple sensor data can extract more accurate degradation features. On the other hand, a new health indicator is constructed to characterize the damage degree of machinery. A suitable health indicator is expected to quantitatively analyze the RUL prediction result. It still needs further research to obtain more accurate and robust RUL prediction results.

Author Contributions: Conceptualization, F.D.; data curation, Y.B.; funding acquisition, Y.L. and S.Y.; methodology, F.D.; writing–original draft, F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Yang S. OF FUNDER, grant number 2020YFB2007700, 11790282, 12032017; This research was funded by Deng F. OF FUNDER, grant number 11802184, E2019210049, A202101017; This research was funded by Liu Y. OF FUNDER, grant number 20310803D, A2020210028.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Available at: https://www.femto-st.fr/en/Research-departments/AS2M/Research-groups/PHM.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qiao, W.; Zhang, P.; Chow, M.-Y. Condition Monitoring, Diagnosis, Prognosis, and Health Management for Wind Energy Conversion Systems. *IEEE Trans. Ind. Electron.* **2015**, *62*, 6533–6535. [CrossRef]
2. Huang, W.; Cheng, J.; Yang, Y.; Guo, G. An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis. *Neurocomputing* **2019**, *359*, 77–92. [CrossRef]
3. Liao, L.; Köttig, F. Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Trans. Reliab.* **2014**, *63*, 191–207. [CrossRef]
4. Gebraeel, N.; Lawley, M.; Liu, R.; Parmeshwaran, V. Residual Life Predictions from Vibration-Based Degradation Signals: A Neural Network Approach. *IEEE Trans. Ind. Electron.* **2004**, *51*, 694–700. [CrossRef]
5. Benkedjouh, T.; Medjaher, K.; Zerhouni, N.; Rechak, S. Remaining useful life estimationbased on nonlinear feature reduction and support vector regression. *Eng. Appl. Artif. Intell.* **2013**, *26*, 1751–1760. [CrossRef]
6. Moosavian, A.; Ahmadi, H.; Tabatabaeefar, A.; Sakhaei, B. An appropriate procedure for detection of journal-bearing fault using power spectral density, k-nearest neighbor and support vector machine. *Int. J. Smart Sens. Intell. Syst.* **2017**, *5*, 685–700. [CrossRef]
7. Dong, M.; He, D. A segmental hidden semi-Markov model (HSMM)-based diagnosticsand prognostics framework and methodology. *Mech. Syst. Signal Process.* **2007**, *21*, 2248–2266. [CrossRef]
8. Liu, Z.; Zuo, M.J.; Qin, Y. Remaining useful life prediction of rolling element bearings based on health stateassessment. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2016**, *230*, 314–330. [CrossRef]
9. Zheng, Y. Remaining useful life estimation of bearings utilizing the k-nearest neighbor classifier and Gaussian process regression. *IPPTA Q. J. Indian Pulp Pap. Tech. Assoc.* **2018**, *30*, 661–671.
10. Ali, J.B.; Chebel-Morello, B.; Saidi, L.; Malinowski, S.; Fnaiech, F. Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network. *Mech. Syst. Signal Process.* **2015**, *56–57*, 150–172.
11. Guo, L.; Lei, Y.; Li, N.; Yan, T.; Li, N. Machinery health indicator construction based on convolutional neural networks considering trend burr. *Neurocomputing* **2018**, *292*, 142–150. [CrossRef]
12. Jiang, P.; Hu, Z.; Liu, J.; Yu, S.; Wu, F. Fault Diagnosis Based on Chemical Sensor Data with an Active Deep Neural Network. *Sensors* **2016**, *16*, 1695. [CrossRef] [PubMed]
13. Deutsch, J.; He, D. Using Deep Learning-Based Approach to Predict Remaining Useful Life of Rotating Components. *IEEE Trans. Syst. Man Cybern. Syst.* **2017**, *48*, 11–20. [CrossRef]
14. Hinchi, A.Z.; Tkiouat, M. Rolling element bearing remaining useful life estimation based on a convolutional long-short-term memory network. *Procedia Comput. Sci.* **2018**, *127*, 123–132. [CrossRef]
15. Chu, C.H.; Lee, C.J.; Yeh, H.Y. Developing Deep Survival Model for Remaining Useful Life Estimation Based on Convolutional and Long Short-Term Memory Neural Networks. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8814658. [CrossRef]
16. Deng, F.; Ding, H.; Yang, S.; Hao, R. An improved deep residual network with multiscale feature fusion for rotating machinery fault diagnosis. *Meas. Sci. Technol.* **2020**, *32*, 024002. [CrossRef]
17. Ren, L.; Sun, Y.; Wang, H.; Zhang, L. Prediction of bearing remaining useful life with deep convolution neuralnetwork. *IEEE Access* **2018**, *6*, 13041–13049. [CrossRef]
18. Li, X.; Zhang, W.; Ding, Q. Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliab. Eng. Syst. Saf.* **2019**, *182*, 208–218. [CrossRef]
19. Li, H.; Zhao, W.; Zhang, Y.; Zio, E. Remaining useful life prediction using multi-scale deep convolutional neural network. *Appl. Soft Comput.* **2020**, *89*, 106113. [CrossRef]
20. Li, X.; Ding, Q.; Sun, J.-Q. Remaining useful life estimation in prognostics using deep convolutionneural networks. *Reliab. Eng. Syst. Saf.* **2018**, *172*, 1–11. [CrossRef]
21. Zhang, W.; Li, C.; Peng, G.; Chen, Y.; Zhang, Z. A deep convolutional neural network with new trainingmethods for bearing fault diagnosis under noisy environmentand different working load. *Mech. Syst. Sig. Process.* **2018**, *100*, 439–453. [CrossRef]

22. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2-4 May 2016.
23. Zhang, H.; Zhang, W.; Shen, W.; Li, N.; Chen, Y.; Li, S.; Chen, B.; Guo, S.; Wang, Y. Automatic segmentation of the cardiac MR images based on nested fully convolutional dense network with dilated convolution. *Biomed. Signal Process. Control* **2021**, *68*, 102684. [CrossRef]
24. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi- Supervised Semantic Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–24 June 2018.
25. Vorugunti, C.S.; Pulabaigari, V.; Mukherjee, P.; Sharma, A. DeepFuseOSV: Online signature verification using hybrid feature fusion and depthwise separable convolution neural network architecture. *IET Biom.* **2020**, *9*, 259–268. [CrossRef]
26. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
27. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
28. Jia, F.; Lei, Y.; Lu, N.; Xing, S. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Syst. Sig. Process* **2018**, *110*, 349–367. [CrossRef]
29. Wang, B.; Lei, Y.; Li, N.; Yan, T. Deep separable convolutional network for remaining useful life prediction of machinery. *Mech. Syst. Signal Process.* **2019**, *134*, 106330. [CrossRef]
30. Nectoux, P.; Gouriveau, R.; Medjaher, K.; Ramasso, E.; Chebel-Morello, B.; Zerhouni, N.; Varnier, C. PRONOSTIA: An experimental platform for bearings accelerated life test. In Proceedings of the IEEE International Conference on Prognostics and Health Management, Denver, CO, USA, 18 June 2012.
31. IEEE PHM 2012 Data Challenge. Available online: http://www.femto-st.fr/f/d/IEEEPHM2012-Challenge-Details.pdf (accessed on 24 October 2021).
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning forimage recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
33. Zhao, R.; Yan, R.; Wang, J.; Mao, K. Learning to monitor machine health with convolutional bi-directional LSTM networks. *Sensors* **2017**, *17*, 273. [CrossRef]